Science & Justice 61 (2021) 299–309

Contents lists available at ScienceDirect

Science & Justice

journal homepage: www.elsevier.com/locate/scijus

Professional Commentary

# Consensus on validation of forensic voice comparison☆

Geoffrey Stewart Morrison [a,b,*], Ewald Enzinger [a], Vincent Hughes [c], Michael Jessen [d], Didier Meuwly [e], Cedric Neumann [f], S. Planting [g], William C. Thompson [h], David van der Vloed [e], Rolf J.F. Ypma [e,a], Cuiling Zhang [i,j,a], A. Anonymous [k], B. Anonymous [k]

[a] Forensic Data Science Laboratory & Forensic Speech Science Laboratory, Department of Computer Science & Aston Institute for Forensic Linguistics, Aston University, Birmingham B4 7ET, UK
[b] Forensic Evaluation Ltd, Birmingham, UK
[c] Department of Language and Linguistic Science, University of York, Heslington, York YO10 5DD, UK
[d] Bundeskriminalamt, Forensic Science Institute, Department of Language and Audio, D-65173 Wiesbaden, Germany
[e] Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, The Hague, The Netherlands
[f] Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA
[g] Public Prosecutor's Office East-Netherlands, Eusebiusbinnensingel 28, 6811 BX Arnhem, The Netherlands
[h] Department of Criminology, Law & Society, University of California, Irvine, CA 92697, USA
[i] School of Criminal Investigation, Southwest University of Political Science & Law, Chongqing, China
[j] Chongqing Institutes of Higher Education Key Forensic Science Laboratory, Chongqing, China
[k] Affiliation witheld

ARTICLE INFO

*Keywords:*
Validation
Likelihood ratio
Guidance
Forensic science
Forensic voice comparison
Admissibility

ABSTRACT

Since the 1960s, there have been calls for forensic voice comparison to be empirically validated under casework conditions. Since around 2000, there have been an increasing number of researchers and practitioners who conduct forensic-voice-comparison research and casework within the likelihood-ratio framework. In recent years, this community of researchers and practitioners has made substantial progress toward validation under casework conditions becoming a standard part of practice: Procedures for conducting validation have been developed, along with graphics and metrics for representing the results, and an increasing number of papers are being published that include empirical validation of forensic-voice-comparison systems under conditions reflecting casework conditions. An outstanding question, however, is: In the context of a case, given the results of an empirical validation of a forensic-voice-comparison system, how can one decide whether the system is good enough for its output to be used in court? This paper provides a statement of consensus developed in response to this question. Contributors included individuals who had knowledge and experience of validating forensic-voice-comparison systems in research and/or casework contexts, and individuals who had actually presented validation results to courts. They also included individuals who could bring a legal perspective on these matters, and individuals with knowledge and experience of validation in forensic science more broadly. We provide recommendations on what practitioners should do when conducting evaluations and validations, and what they should present to the court. Although our focus is explicitly on forensic voice comparison, we hope that this contribution will be of interest to an audience concerned with validation in forensic science more broadly. Although not written specifically for a legal audience, we hope that this contribution will still be of interest to lawyers.

☆ This consensus is also supported by: Fanny Carlström Plaza (Swedish National Forensic Centre); Joaquín González-Rodríguez (Universidad Autónoma de Madrid); Daniel Ramos (Universidad Autónoma de Madrid); Paul Roberts (University of Nottingham / China University of Political Science and Law); Phil Rose (Forensic Speech Science Consultant); Yosef Solewicz (Israel National Police); P. Vergeer (Netherlands Forensic Institute).

* Corresponding author.
*E-mail address:* geoff-morrison@forensic-evaluation.net (G.S. Morrison).

https://doi.org/10.1016/j.scijus.2021.02.002
Received 9 September 2020; Received in revised form 13 February 2021; Accepted 21 February 2021
Available online 6 March 2021

## 1. Introduction

Since the 1960s, there have been calls for forensic voice comparison[1] to be empirically validated under casework conditions (for a review, see [1]). Since around 2000, there have been an increasing number of researchers and practitioners who conduct forensic-voice-comparison research and casework within the likelihood-ratio framework.[2] In recent years, this community of researchers and practitioners has made substantial progress toward validation under casework conditions becoming a standard part of practice:

- Procedures for conducting validation have been developed, along with graphics and metrics for representing the results, e.g., Tippett plots [2] and the log-likelihood-ratio cost ($C_{\text{llr}}$) [3].[3]
- An increasing number of papers are being published that include empirical validation of forensic-voice-comparison systems under conditions reflecting casework conditions, e.g., [4–19].

An outstanding question, however, is:

- In the context of a case, given the results of an empirical validation of a forensic-voice-comparison system, how can one decide whether the system is good enough for its output to be used in court?

In some jurisdictions and situations, this question may be related to a formal admissibility decision made by a judge. In other jurisdictions or situations, it may be related to what weight, if any, the trier of fact gives to the output of a forensic-voice-comparison system. It could also be related to whether a forensic practitioner decides to proceed with evaluation[4] of the questioned- and known-speaker recordings from a case, or to whether a lawyer decides to tender the results of a forensic voice comparison as evidence in court, or to whether a lawyer decides to use those results in pre-trial negotiations.

Our intent is to address this question from a scientific perspective, rather than a legal perspective. Our intent is to describe the consensus as to what is generally accepted within the relevant scientific community.[5] To this end, §2 below presents our statement of consensus with respect to validation of forensic-voice-comparison systems. The methodology by which we arrived at this statement of consensus is described in Appendix D.

Since the statement of consensus is not a national or international standard, we provide recommendations only, and not requirements or permissions. For ease of reference, the paragraphs in §2 are numbered. Some paragraphs state recommendations, but others are explanatory only. Sentences containing "should" state recommendations. In order to keep the statement of consensus succinct, background information is provided in appendices. The introduction and appendices are informational only, and do not form part of the statement of consensus.

The primary intended audience is forensic practitioners who conduct forensic-voice-comparison evaluations and validations or who critique forensic-voice-comparison evaluation and validation reports prepared by others. We provide recommendations on what practitioners should do when conducting evaluations and validations, and what they should present to the court. Although our focus is explicitly on forensic voice comparison, we hope that this contribution will be of interest to an audience concerned with validation in forensic science more broadly. Although not written specifically for a legal audience, we hope that this contribution will still be of interest to lawyers.

## 2. Statement of consensus with respect to validation of forensic-voice-comparison systems

### 2.1. Scope

2.1.1. This statement of consensus addresses validation for the purpose of demonstrating whether, in the context of specific cases, forensic-voice-comparison systems are (or are not) good enough for their output to be used in court. Validation for system-development purposes and validation for investigative applications are out of scope.

2.1.2 This statement of consensus addresses scientific matters that could have a bearing on legal decisions, but it does not address legal matters directly.

2.1.3. This statement of consensus applies to validation of forensic-voice-comparison systems that are based on relevant data, quantitative measurements, and statistical models, and that output numeric likelihood ratios. Some recommendations assume that these systems are implementations of human-supervised-automatic approaches.

### 2.2. Calculating a likelihood ratio: Propositions, relevant population, and conditions

2.2.1. In the context of forensic inference, a likelihood ratio provides an answer to a specific question formed by two mutually exclusive propositions. Those propositions should include specification of what constitutes the relevant population for the case under consideration. See Appendix A.

2.2.2. The forensic practitioner should clearly communicate to the court what propositions they have adopted, including clearly describing what they have adopted as the relevant population. This is a prerequisite for the court to be able to understand the question addressed by the forensic practitioner and consider whether it is a relevant and appropriate question. Understanding the question is also a prerequisite for understanding the answer.

2.2.3. The relevant population and the conditions of questioned-speaker and known-speaker recordings can vary from case to case and there can be a mismatch between the conditions of the questioned-speaker recording and the known-speaker recording in a case.[6] For a brief explanation of recording conditions, see Appendix B.

2.2.4. The forensic practitioner should clearly communicate to the court what the forensic practitioner understands the conditions of the questioned-speaker and known-speaker recordings to be.

### 2.3. Calculating a likelihood ratio: Calibration

2.3.1. In order for the forensic-voice-comparison system to answer the specific question formed by the propositions in the case, the

[1] "Forensic voice comparison" is also known as "forensic speaker recognition", "forensic speaker identification", and "forensic speaker comparison".

[2] The likelihood-ratio framework is described in Appendix A of the present paper.

[3] Descriptions of Tippett plots and $C_{\text{llr}}$ are provided in Appendix C of the present paper.

[4] "Evaluation" comprises both "analysis" and "interpretation". "Analysis" is the process of extracting information from the objects of interest in the case (in present context, the questioned- and known-speaker recordings). "Interpretation" is the process of drawing inferences from that information (in present context, calculating a likelihood ratio that addresses relevant propositions for the case).

[5] In the United States, "general acceptance within the relevant scientific community" is the admissibility criterion in Frye v United States, 293F 1013 (DCCir 1923), and is an admissibility criterion in Daubert v Merrell Dow Pharmaceuticals, 509 US 579 (1993).

[6] For simplicity, the present document is written in a manner that assumes a single questioned-speaker recording and a single known-speaker recording.

output of the system should be well calibrated. For an explanation of what constitutes a well calibrated system, see §C.1 in Appendix C.

2.3.2 A forensic-voice-comparison system should be calibrated using a statistical model that forms the final stage of the system (hereinafter the "calibration model").[7]

2.3.3. Data used for training[8] the calibration model (hereinafter "calibration data") should be sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case, that, when the system is used to compare the questioned- and known-speaker recordings, the resulting likelihood ratio will be a reasonable answer to the question posed by the propositions.[9]

### 2.4. Validation procedures

2.4.1. In order to validate a forensic-voice-comparison system, pairs of recordings should be input to the system and the likelihood-ratio output corresponding to each pair obtained. (Hereinafter these pairs of recordings are collectively referred to as "validation data".)

2.4.2. Some pairs of recordings should be same-speaker pairs (both members of the pair were produced by the same speaker), and other pairs of recordings should be different-speaker pairs (each member of the pair was produced by a different speaker). The system being validated should not have access to information as to the true status of each pair, i.e., whether it is a same-speaker or a different-speaker pair.[10]

2.4.3. The result is a set of same-speaker likelihood-ratio values (values calculated when it is known that the input was a same-speaker pair), and a set of different-speaker likelihood-ratio values (values calculated when it is known that the input was a different-speaker pair). The performance of the system is then assessed by comparing the likelihood-ratio values output by the system with the truth as to whether they resulted from same-speaker or different-speaker comparisons. If the performance of the system is good, same-speaker likelihood-ratio values will be large and different-speaker likelihood-ratio values will be small.

### 2.5. Validation data

2.5.1. For each pair of recordings in the validation data, one member of the pair should have conditions that reflect those of the questioned-speaker recording in the case, and the other member of the pair should have conditions that reflect those of the known-speaker recording in the case.

2.5.2. Validation data should be sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case, that the results of validating the system using those data will be informative as to the expected performance of the system when it is applied in the case.[11]

2.5.3. One of the criteria for the validation data to be sufficient is that the number of speakers included be sufficient. Because of sampling variability, small validation sets can give results that are not representative of the case conditions.

2.5.4. Data used for validation should not include recordings of the same speakers as were used for any part of system training (including training the calibration model). Either separate data sets should be used or appropriate cross-validation should be used.[12]

2.5.5. The forensic-voice-comparison system will ultimately be used to calculate a likelihood-ratio value for a comparison of a pair of recordings it has not been trained on, the questioned-speaker and known-speaker recordings in the case. Validating using recordings of the same speakers as were used for training will give overly optimistic results.[13]

### 2.6. Decision as to whether calibration and validation data are sufficient

2.6.1. The decision as to whether the calibration data and the validation data are sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case will be the result of a subjective judgment made by the forensic practitioner.

2.6.2. A system in which the conclusion is the direct result of a subjective judgment is susceptible to cognitive bias. By restricting subjective judgments to the earliest steps in the interpretive process, however, susceptibility to cognitive bias is substantially reduced.[14]

2.6.3. If relevant research results are available, the decision as to whether the calibration and validation data are sufficient should be informed by research on the effects of changes in data sets on the performance of the forensic-voice-comparison system (or the type of forensic-voice-comparison system) that the practitioner is using.

2.6.4. If relevant metrics are available, the decision as to whether the calibration and validation data are sufficient should be informed by the use of quantitative metrics of the degree of mismatch between case recordings versus calibration and validation recordings.[15]

[7] For an introduction to calibration of forensic-evaluation systems that output likelihood ratios, see [20]. Note that the scores that are to be calibrated in forensic voice comparison are scores that take account of both similarity and typicality. These scores are uncalibrated likelihood ratios. They are not similarity-only scores. See discussion in [21–23].

[8] In the present document, "training" a statistical model is intended to cover both training a statistical model from scratch using only the case-specific data, and, if applicable, using the case-specific data to adapt an existing model.

[9] §2.6 discusses the decision as to whether calibration data are sufficient.

[10] This recommendation is not intended to exclude the use of appropriate cross-validation, see note 12.

[11] §2.6 discusses the decision as to whether validation data are sufficient.

[12] If cross-validation is used, leave-one-speaker-out / leave-two-speakers-out cross-validation should be used for training the calibration model. This minimizes the differences between the data used to train the calibration model in each cross-validation loop. It also minimizes the differences between the calibration models in the cross-validation loops and the calibration model that is used to calibrate the questioned-speaker-versus-known-speaker score. The latter model should be trained on the full set of calibration data. In a cross-validation loop in which the score to be calibrated is a same-speaker score, e.g., a recording of speaker *A* compared to another recording of speaker *A*, all scores that resulted from comparisons in which one or both members of the pair was a recording of speaker *A* should be excluded from the data used to train the calibration model (leave-one-speaker-out). In a cross-validation loop in which the score to be calibrated is a different-speaker score, e.g., a recording of speaker *A* compared to a recording of speaker *B*, all scores that resulted from comparisons in which one or both members of the pair was a recording of speaker *A* or a recording of speaker *B* should be excluded from the data used to train the calibration model (leave-two-speakers-out).

[13] "In academic settings, we usually do have access to the test set, but we should not use it for model fitting or model selection, otherwise we will get an unrealistically optimistic estimate of performance of our method. This is one of the 'golden rules' of machine learning research." ([24] p. 23 n. 11)

[14] Cognitive bias in forensic science is of increasing concern – for reviews see [25–28].

[15] In the present document, "mismatch" refers to differences in population or condition, not to differences between individual speakers. Simple metrics of degree of mismatch could be based on properties such as signal-to-noise ratio, net duration of speech, signal bandwidth, or compression artifacts.

2.6.5. The forensic practitioner should clearly communicate to the court that the decision as to whether the calibration and validation data are sufficient is based on subjective judgment.

2.6.6. The forensic practitioner should communicate to the court the basis for this decision, including referencing any research reports consulted and providing the values of any degree-of-mismatch metrics that contributed to the decision.

2.6.7 The forensic practitioner should communicate to the court a clear description of the calibration data and the validation data used.

2.6.8. A description of the calibration and validation data is a prerequisite for a second forensic practitioner to be able to conduct an independent review so as to be able to opine on whether the data are sufficient.

2.6.9. A description of the calibration and validation data is also a prerequisite for the court to be able to decide to either accept or reject the first forensic practitioner's decision about the sufficiency of the data.

### *2.7. Anticipatory and case-by-case validation*

2.7.1 Validation can be conducted in anticipatory mode or in case-by-case mode.

2.7.2. In anticipatory mode, a forensic-voice-comparison system is validated under sets of conditions that are expected to occur in future casework. When beginning a new case, the practitioner should make a judgment as to whether the relevant population and conditions for the case are sufficiently similar to those in an existing validation report that that report can be used in conjunction with the case.

2.7.3. Alternatively, if the relevant population and the conditions of the new case are not sufficiently similar to any single existing validation report, but the practitioner judges that performance for the case can be interpolated or extrapolated from multiple existing validation reports, those validation reports, accompanied by an explanation of the interpolation or extrapolation, should be used in conjunction with the case.

2.7.4. If the practitioner judges that no existing validation reports are based on data that are sufficiently similar to the relevant population and conditions for the new case, then (if the practitioner is to proceed with the case) the practitioner should obtain data that they judge to be sufficient and conduct a new validation for that case using those data. This is case-by-case validation.

2.7.5. If the practitioner cannot access or generate one or more validation reports that they judge to be based on data that are sufficiently similar to the relevant population and conditions of the case under consideration, the practitioner should terminate their evaluation. They should not proceed to use the forensic-voice-comparison system to compare the questioned-speaker and known-speaker recordings in the case.

### *2.8. Presenting validation results*

2.8.1. Validation results should be presented in a validation report. The validation report should be provided to the court. (The validation report could be provided as part of the casework report.)

2.8.2. Validation results should be presented using graphics and metrics that are appropriate for representing the performance of systems that output numeric likelihood ratios. An appropriate graphic is a Tippett plot, and an appropriate metric is the log-likelihood ratio cost ($C_{llr}$), see Appendix C §C.1 and §C.2 respectively.

*2.8.3.* $C_{llr}$ should be calculated and included in the validation report, and a Tippett plot should be drawn and included in the validation report.

2.8.4. If the $C_{llr}$ value is greater than 1, the system is not well calibrated (which could potentially be remedied by adding a calibration model to the system).

2.8.5. Even if $C_{llr}$ is less than 1, however, this does not guarantee that the system is well calibrated. Miscalibration could still be apparent in the Tippett plot (see §C.1 in Appendix C).

2.8.6. The practitioner should communicate to the court whether the practitioner observes any indications of miscalibration in the validation results, i.e., a $C_{llr}$ value greater than 1 and/or bias apparent in the Tippett plot.

2.8.7. If the validation results are not well calibrated, the practitioner should terminate their evaluation. They should not proceed to use the forensic-voice-comparison system to compare the questioned-speaker and known-speaker recordings in the case.

### *2.9. Relationship between conditions and performance*

2.9.1 The demonstrated performance of a forensic-voice-comparison system depends on:

a) the properties of the system, including the calibration model; and

b) the properties of the validation data.

2.9.2. Given two systems, under a particular set of conditions the performance of the first system could be better than the second, but under a different set of conditions the performance of the second system could be better than the first.

2.9.3. Validating under some conditions could result in poorer performance than validating under other conditions, e.g., conditions involving shorter recordings and greater background noise would be expected to lead to poorer performance. It could also be that validating using samples from one population results in poorer performance than validating using samples from another population. As conditions become more challenging, system performance will become poorer.

### *2.10. Validation threshold for* $C_{llr}$

2.10.1. Assuming a system is well calibrated, what constitutes poorer performance are likelihood-ratio values that are on average closer to 1 – likelihood-ratio values resulting from both same-speaker pairs and different-speaker pairs will on average be closer to 1 than they would be under less challenging conditions.

2.10.2. Dismissing a likelihood-ratio value because it is relatively close to 1 is a form of the "defense attorney's fallacy" [29]: If there are multiple pieces of evidence in a case and they all give likelihood-ratio values that are relatively close to 1 but that all point in the same

direction (they are all above 1, or they are all below 1), then the combined strength of all the evidence could be substantial (when all the likelihood-ratio values are multiplied together, the combined strength of evidence could be far from 1).[16] A single likelihood-ratio value should not, therefore, be dismissed just because it is relatively close to 1.

2.10.3. A well-calibrated system that has poor performance will output likelihood-ratio values that tend to be relatively close to 1. By extension of the argument in the previous paragraph, a system should not be rejected just because the likelihood-ratio values it outputs tend to be relatively close to 1.

2.10.4. A well-calibrated system that has poor performance will have a relatively high $C_{llr}$ value; however, assuming the validation data are sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case, as long as $C_{llr}$ is less than 1 the system is providing useful information. If $C_{llr}$ equals 1, then on average the system is no better than a system that always responds with a likelihood-ratio value of 1 irrespective of the input. For the latter system the posterior odds would always equal the prior odds, hence the system would never provide any useful information.

2.10.5. As explained above, as long as its $C_{llr}$ is less than 1, a system is providing useful information. Use of a forensic-voice-comparison system should not, therefore, be rejected just because its $C_{llr}$ value is high. The only logically justified validation-threshold value for $C_{llr}$ is 1.

2.10.6. The practitioner should communicate to the court whether, in the practitioner's opinion, the system is providing useful information.

### *2.11. Decision as to whether the likelihood-ratio value for the comparison of the questioned-speaker and known-speaker recordings is supported by the validation results*

2.11.1. A Tippett plot displays all the likelihood-ratio values generated using the validation data and gives an indication of the range of likelihood-ratio values that could be expected given the relevant population for the case and the conditions of the questioned-speaker and known-speaker recordings in the case. This allows for a check of whether the likelihood-ratio value calculated for the comparison of the questioned-speaker and known-speaker recordings is supported by the validation results.

2.11.2. For example, if the Tippett plot included likelihood-ratio values in the range 1/1000 to 100, but the likelihood-ratio value calculated for the comparison of the questioned-speaker and known-speaker recordings was 10,000, then this would be suspicious. A value so far beyond the range of values obtained in the validation results would likely be due to a mistake, e.g., it could be that calibration data and/or the validation data do not actually represent the population to which the questioned speaker belongs, or it could be that they do not actually reflect the conditions of the questioned-speaker and known-speaker recordings.

2.11.3. A likelihood-ratio value calculated for the comparison of the questioned-speaker and known-speaker recordings that is within the range shown in the Tippett plot would unambiguously be supported by the validation results, and a value just beyond the range would be reasonable.

2.11.4. The forensic practitioner should communicate to the court whether, in the forensic practitioner's opinion, the likelihood-ratio value calculated for the comparison of the questioned-speaker and known-speaker recordings is supported by the validation results.

### *2.12. Summary of key points*

2.12.1. The forensic practitioner should communicate to the court what propositions the forensic practitioner has adopted for the case, including what they have adopted as the relevant population.

2.12.2. The forensic practitioner should communicate to the court what the forensic practitioner understands the conditions of the questioned-speaker and known-speaker recordings to be.

2.12.3 The forensic-voice-comparison system should be well calibrated.

2.12.4. Validation data should be representative of the relevant population for the case, and reflective of the conditions of the questioned-speaker and known-speaker recordings in the case.

2.12.5. The forensic practitioner's decision as to whether the validation data are sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case, will be a subjective judgment.

2.12.6. Validation results should be presented as a Tippett plot and a $C_{llr}$ value. These should be examined for signs of miscalibration.

2.12.7. The validation threshold (acceptance criterion) for $C_{llr}$ should be 1. As long as $C_{llr}$ is less than 1, the system is providing useful information.

2.12.8. To decide whether the likelihood-ratio value calculated for the comparison of the questioned-speaker and known-speaker recordings is supported by the validation results, it should be compared with the values shown in the Tippett plot.

## 3. Disclaimer

The contents of this document represent a consensus reached among the authors, and agreed to by the supporters. This consensus does not necessarily reflect the policies or positions of any organizations with which the authors or supporters are affiliated.

## Declaration of Competing Interest

none

*Acknowledgements*

The initial meeting, Morrison's contribution, and open access publication of this document were supported by a Research England Expanding Excellence in England grant awarded to the Aston Institute for Forensic Linguistics. The initial meeting was also supported by the Netherlands Forensic Institute.

Thompson's contribution was supported by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through a Cooperative Agreement with the National Institute of Standards and Technology (NIST) (Cooperative Agreement No. 70NANB20H019).

Zhang's contribution was supported by the National Social Science Foundation of China Key Program (Grant No. 16AYY015), and Chongqing Social Enterprise and People's Livelihood Guarantee Scientific and Technological Innovation Special Research and Development

[16] The latter calculation is provided for explanatory purposes. The general point is valid, but, for simple multiplication to provide an accurate result, all the likelihood ratios would have to be based on the same pair of propositions and they would have to be statistically independent, e.g., because they are based on unrelated types of evidence.

Key Project (cstc2017shms- zdyfX0060).

## Appendix A. The likelihood-ratio framework

### A.1. Questions and answers

In the context of forensic interpretation, a likelihood ratio provides the answer to a specific two-part question, for example:[17]

(a) What is the likelihood of obtaining the observed properties of the voices of interest on the questioned- and known-speaker recordings if they were both produced by the same speaker, a speaker selected at random from the relevant population?
versus
(b) What is the likelihood of obtaining the observed properties of the voices of interest on the questioned- and known-speaker recordings if they were each produced by a different speaker, each speaker selected at random from the relevant population?

Part (a) is a question corresponding to the proposition that the voices of interest on two or more recordings were produced by the same speaker (the *same-speaker proposition*), and part (b) is a question corresponding to the proposition that the voices of interest on two or more recordings were produced by different speakers (the *different-speaker proposition*). The same-speaker and different-speaker propositions are also known as the same-speaker and different-speaker hypotheses.

The answer to part (a) quantifies the *similarity* between the voices of interest on the questioned- and known-speaker recordings, and the answer to part (b) quantifies the *typicality* of the voices of interest on the questioned- and known-speaker recordings. Typicality is defined with respect to the relevant population.

The likelihood ratio is the result of dividing the answer to part (a) by the answer to part (b).

### A.2. Relevant population

The relevant population is the population from which the questioned speaker is hypothesized to have come if the questioned speaker were not the known speaker. Depending on the circumstances of the case, the relevant population could be a very large group of speakers, a small group of speakers, or a single speaker who is not the known speaker. Specification of what constitutes the relevant population is a key part of the specification of the propositions (particularly the different-speaker proposition), and hence is a key part of the specification of the question that is being answered.

### A.3. The meaning of a likelihood ratio

The following text is provided to explain the meaning of a likelihood ratio. It is not intended as an attempt to instruct a court of law as to how to reason on legal matters.

The likelihood ratio calculated for the comparison of the questioned- and known-speaker recordings constitutes the forensic practitioner's conclusion as to the strength of the evidence. Logically, the likelihood ratio quantifies the amount by which the decision maker should update their belief with respect to the probability that the same-speaker proposition is true versus the probability that the different-speaker proposition is true. This is formally expressed in Equation 1, which is a form of Bayes' Theorem (it is the "odds form" of Bayes' Theorem).

*prior odds* × *likelihood ratio* = *posterior odds*

$$\frac{p(H_s)}{p(H_d)} \times \frac{f(E|H_s)}{f(E|H_d)} = \frac{p(H_s|E)}{p(H_d|E)} \tag{1}$$

The *prior odds* quantify the decision maker's belief that the same-speaker proposition is true divided by their belief that the different-speaker proposition is true *before* the forensic practitioner presents their conclusion as to the strength of evidence. The *likelihood ratio* is what the forensic practitioner presents as their strength-of-evidence conclusion. The *posterior odds* quantify the decision maker's belief that the same-speaker proposition is true divided by their belief that the different-speaker proposition is true *after* the forensic practitioner has presented their conclusion as to the strength of evidence. According to Bayes' Theorem, in order to update their beliefs, the decision maker should multiply their prior odds by the likelihood ratio to arrive at their posterior odds.

### A.4. Further reading

General introductions to the likelihood-ratio framework include [31] and [32]. Introductions to the likelihood-ratio framework in the context of forensic voice comparison include [33]. A more-advanced introduction to statistical models used for calculating likelihood ratios in human-supervised-automatic approaches to forensic voice comparison is provided in [34].

## Appendix B. Recording conditions

The following two paragraphs are based in-part on [35] pp. 76–77. A fuller discussion of intra-speaker variability and recording conditions in speech-, speaker-, and language-recognition tasks is provided in [36].

Variation in the conditions of recordings can be due to speaker intrinsic factors. The way a speaker speaks can vary from occasion to occasion

[17] This is an example of a *common-source* likelihood ratio (see [30] on the distinction between *common-source* and *specific-source* likelihood ratios). Statistical models used in modern human-supervised-automatic forensic-voice-comparison systems calculate common-source likelihood ratios.

because of a variety of factors, including: speaking style due to situation or interlocutor (e.g., formal versus casual); vocal effort (whispering versus shouting being extremes, but moderately increased vocal effort due to background noise or perceived communication difficulty is common in forensic casework); cognitive load; physical stress; emotions; health conditions; and deliberate disguise.

Variation in the conditions of recordings can be due to speaker extrinsic factors. These can include factors such as: different types and degrees of background noise; reverberation; distance to microphone; frequency response of the microphone and other components of the recording system; sampling rate and quantization level for digitization; transmission through communication channels (e.g., landline telephone, mobile telephone, voice-over-internet protocol, radio transmission); and codecs used for transmission or for saving the recording (lossy compression is common for reducing the amount of information transmitted or for reducing the amount of storage space needed). For examples in forensic contexts, and that present the results as Tippett plots and $C_{\text{llr}}$, see [37] and [19].

The duration of the speech of the speaker of interest on each recording is also part of the conditions, as is the time elapsed between when the questioned- and known-speaker recordings were made. The variability between recordings of the same speaker tends to increase as the time interval increases, especially as it extends into several years. For examples in a forensic context, and that present the results as Tippett plots and $C_{\text{llr}}$, see [38] and [39].

## Appendix C. Tippett plots and $C_{\text{llr}}$

### C.1. Tippett plots

Tippett plots were first proposed in [2]. They were named in honor of C.F. Tippett. The idea of plotting likelihood-ratio results as empirical cumulative probability distributions was not new (see, for example, [40]), but the innovation in Tippett plots was to include the empirical cumulative probability distributions of both same-speaker and different-speaker likelihood ratio values on a single plot. An advantage of the empirical cumulative probability distribution over other graphical representations such as histograms or kernel density plots is that it represents the exact values of the output of the system. Tippett plots graphically represent each and every likelihood-ratio output corresponding to each and every input pair. Descriptions of Tippett plots can be found in [41–45].

Fig. 1 and the text of the next paragraph are adapted from [34].

Three example Tippett plots are shown in Fig. 1. The plots are based on artificial data created for illustrative purposes. The $y$-axis values corresponding to the curves rising to the right give the proportion of same-speaker validation results with log-likelihood-ratio values less than or equal to the corresponding value on the $x$ axis. The $y$-axis values corresponding to the curves rising to the left give the proportion of different-speaker validation results with log-likelihood-ratio values greater than or equal to the corresponding value on the $x$ axis. In general, a Tippett plot in which the two curves have greater separation and in which the curves are shallower indicates better performance.

Note that the curves do not extend to a $y$ value of zero as they are representations of the empirical cumulative probability distribution, hence the lowest $y$ value corresponds to $1/N$ where $N$ is the number of same-speaker or different-speaker input pairs (for these illustrative data $N_s = 50$ and $N_d = 200$). Note also that the $x$ values of the curves are not extrapolated beyond the values of the actual validation results obtained.

Tippett plots can reveal problems such as bias in the output. For a perfectly-calibrated system, the likelihood ratios of the likelihood-ratio values that it outputs will be the same as the likelihood-ratio values that it outputs. For a well-calibrated system, they will be approximately the same. Other than because of differences due to sampling variability between calibration and validation data, calibrating the output of an already well-calibrated system will not change that output. There are two basic forms of potential bias in the output: 1. Shift: All the likelihood-ratio values, originating from both different-speaker input pairs and same-speaker input pairs, are either too big or too small. 2. Scaling: All the likelihood-ratio values are either too far away from the neutral value of 1 or too close to the neutral value of 1 (log-likelihood ratio values are too far away from 0 or too close to 0). A system could exhibit bias in the form of both shift and scale. The top panel of Fig. 1 shows a Tippett plot of the output of a system that is not calibrated. The middle panel shows the same output after calibration. The Tippett plot in the top panel exhibits both shift and scale bias: the log-likelihood-ratio values are too high (the curves are too far to the right) and too close to their intersect value (the slopes of the curves are too steep, the intersect value is not close to the neutral value of 0 because of the shift). In contrast, in the middle panel, the log-likelihood-ratio values are centered around 0 and are on-average further from the neutral value of 0 (the intersect of the curves is close to 0 and their slopes are shallower). The bottom panel shows the output of another well-calibrated system that has better performance than the system whose output is shown in the middle panel. In the bottom panel, the slopes are shallower and the intersect lower.

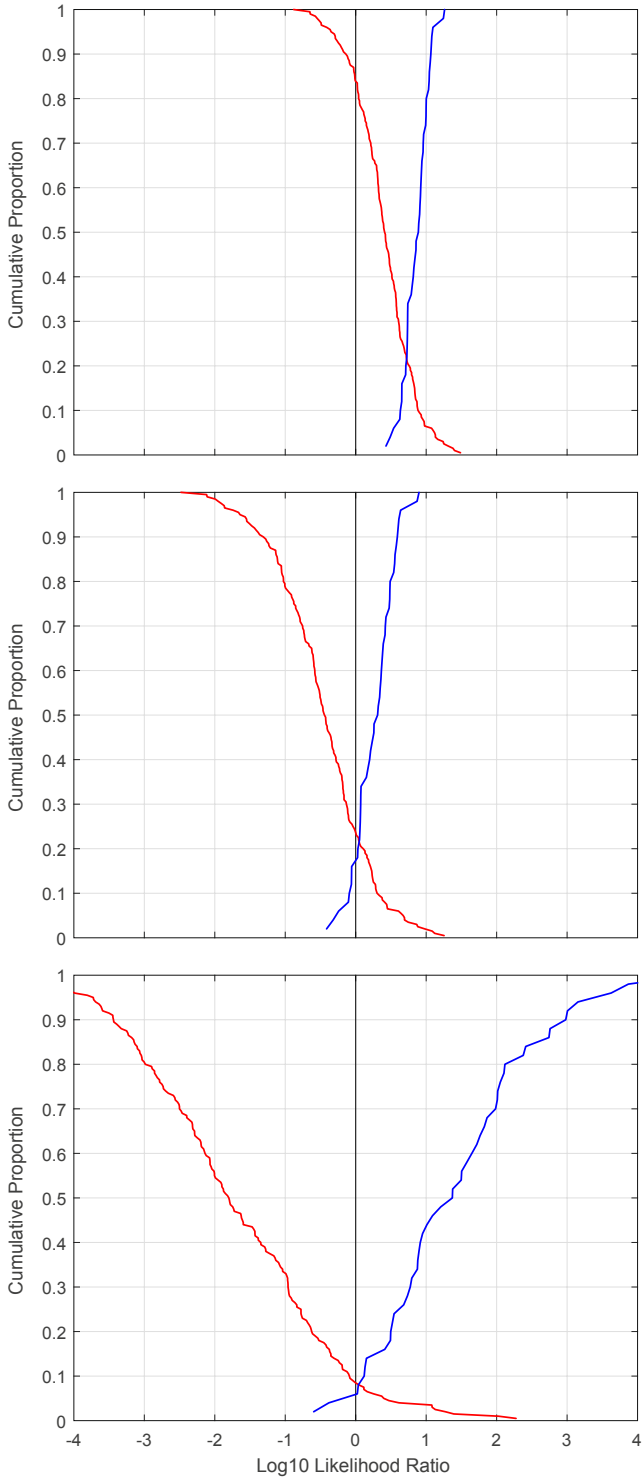


**Fig. 1.** Examples of Tippett plots (see main text for details).

### *C.2. Log-likelihood-ratio cost ($C_{llr}$)*

The log-likelihood-ratio cost ($C_{\text{llr}}$) was first proposed in [3]. It is equivalent to the deviance statistic, assuming equal priors. Descriptions of $C_{\text{llr}}$ can be found in [41,43–47].

Fig. 2 and the text of the next two paragraphs are adapted from [34].

$C_{\text{llr}}$ is calculated using Equation 2, in which $\Lambda_s$ and $\Lambda_d$ are the values of likelihood ratio outputs corresponding to same-speaker and different-speaker inputs respectively, and $N_s$ and $N_d$ are the number of same-speaker and different-speaker inputs respectively.

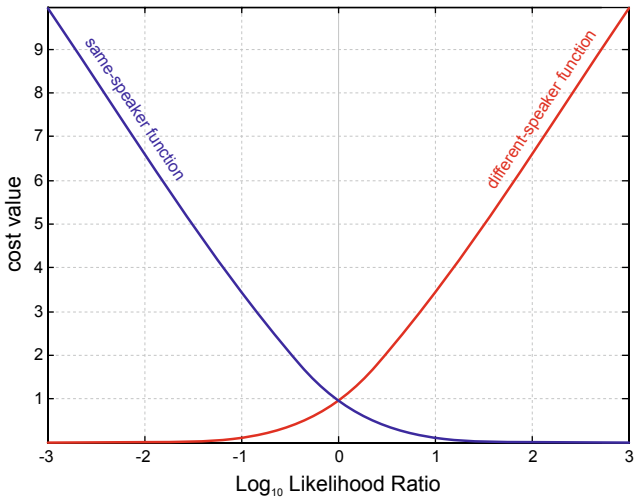


**Fig. 2.** Cost functions for calculating $C_{\text{llr}}$ (see main text for details).

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_s}\sum_{i}^{N_s}\log_2\left(1+\frac{1}{\Lambda_{s_i}}\right)+\frac{1}{N_d}\sum_{j}^{N_d}\log_2\left(1+\Lambda_{d_j}\right)\right) \tag{2}$$

Given a same-speaker input, a good output would be a likelihood-ratio value that is much larger than 1, a less good output would be a value that is only a little larger than 1, a bad output would be a value less than 1, and a worse output would be a value much less than 1. *Mutatis mutandis* for a different-speaker input for which a good output would be a value much less than 1. Fig. 2 plots the cost functions for log-likelihood-ratio outputs corresponding to same-speaker and different-speaker input pairs. These are the functions within Equation 2s left and right summations respectively. If the input is a likelihood ratio from a same-speaker pair and its value is much greater than 1 (its log-likelihood-ratio value is much greater than 0) it receives a small cost value, but if its value is lower it receives a higher cost value. If the input is a likelihood ratio from a different-speaker pair and its value is much less than 1 (its log-likelihood-ratio value is much less than 0) it receives a small cost value, but if its value is higher it receives a higher cost value. $C_{\text{llr}}$ is calculated as the mean of the cost values with the same weight given to the set of same-speaker cost values as to the set of different-speaker cost values.

Smaller $C_{\text{llr}}$ values indicate better performance. $C_{\text{llr}}$ values cannot be less than or equal to 0. For well-calibrated systems, $C_{\text{llr}}$ values lie in the range 0 to approximately 1. A well-calibrated system that performed at the level of chance would have a $C_{\text{llr}}$ value of approximately 1.[18] A $C_{\text{llr}}$ value less than 1 does not necessarily imply that the system is well calibrated; miscalibration may be apparent in the Tippett plot. $C_{\text{llr}}$ values substantially greater than 1 can be produced by uncalibrated or miscalibrated systems.

The $C_{\text{llr}}$ values corresponding to the validation results shown in the Tippett plots of Fig. 1 are 1.068, 0.698, and 0.307 for the top, middle, and bottom panels respectively.

## Appendix D. Methodology

### *D.1. Participants*

In June 2019, invitations to participate in the consensus-development process were extended to 21 individuals. Invitees were individuals who when brought together could be considered representative of the relevant scientific community. They included individuals who had knowledge and experience of validating forensic-voice-comparison systems in research and/or casework contexts, and individuals who had actually presented validation results to courts. They also included individuals who could bring a legal perspective on these matters, and individuals with knowledge and experience of validation in forensic science more broadly.

A two-day meeting was held in September 2019. The meeting was organized and sponsored by the Forensic Speech Science Laboratory of the Aston Institute for Forensic Linguistics and was hosted by the Netherlands Forensic Institute. Not all invitees were able to participate in the meeting. Twelve invitees participated. Eleven attended in-person and one contributed by videoconference. Prior to the meeting, participants were informed of the scope, asked to review relevant literature, and asked to come to the meeting prepared to share their thoughts on the topic.

### *D.2. Scope*

The scope was restricted to validation of forensic-voice-comparison systems that output numeric likelihood ratios. Extensive discussion of other branches of forensic science was out of scope. The scope was also restricted to validation for the purpose of demonstrating whether, in the context of specific cases, a forensic voice-comparison-system is (or is not) good enough for its output to be used in court. Validation for system development and validation for investigative applications were out of scope.

[18] A system that gave no useful information and always responded with a likelihood ratio of 1, irrespective of the input, would have a $C_{\text{llr}}$ value of 1.

### D.3. Process during the initial meeting – Verbal discussion and summarization

On the first day of the meeting, after a verbal recap of the topic and scope by the moderator (the first author of the present paper), each attendee was asked to share their thoughts on the topic. After each attendee had done this, the remainder of the day was dedicated to lightly-moderated discussion. Moderation was kept to a minimum so as not to stifle discussion of a broad range of ideas. The discussion was only redirected if it had clearly gone outside the defined scope and did not look like it was naturally moving back within scope. Participants who had not spoken much were specifically invited to share their thoughts.

On the second day, the moderator attempted to summarize what appeared to have emerged as the consensus, and asked participants to indicate if they were indeed in agreement and to help modify and refine that summary so that it reflected the consensus. In addition to verbal discussion, a written summary of points of agreement was produced. The written summary was in note form.

### D.4. Process after initial meeting: Stage 1 – drafting, verbal discussion, and revision

After the initial meeting, based on the notes as to the consensus reached during the meeting, the editor (the first author of the present paper) produced a first draft of the present paper. This draft was distributed to those who had participated in the meeting. Participants were asked to consider the concepts (rather than the exact wording), and to provide their input during three videoconferences that were held in January, February, and March 2020. During the videoconferences, notes on participants' input were made and a consensus as to how to proceed was agreed.

After the three videoconferences, based on the notes and consensus reached during the discussion, the editor revised the existing draft.

### D.5. Process after initial meeting: Stage 2 – Written comments and proposals for change

The draft resulting from Stage 1 was distributed to participants. We then followed a formal commenting process similar to that used by standards development organizations such as the International Organization for Standardization (ISO): Participants were asked to complete and submit comment sheets in which they had to identify relevant sections of the document, comment on those sections, and make concrete proposals for changes (each comment had to justify the reason for an associated proposed change). We then met via videoconference to discuss the submitted comments and decide which proposals to adopt. The editor then implemented the agreed changes. The cycle of submission of written comments and videoconference occurred five times during April through August 2020 (some cycles required two videoconferences to cover all the comments). For the first three rounds, comments and proposals were restricted to the actual statement of consensus (§2 of the present paper).

A final version of the present paper was produced, and those who had participated were invited to include their names in the published list of authors. For reasons unrelated to the content of the final version, 2 participants did not include their names in the published list of authors. Those who had been invited to the original meeting, but had been unable to attend, were also invited to add their names as supporters of the consensus.[19]

The manuscript was submitted for publication in September 2020. Comments from a single reviewer were received in February 2021. The reviewer's only requested change was the removal of one clause and an associated footnote. This material was explanatory only, it did not include a recommendation. The clause and footnote were deleted, additional proofreading corrections were made, and the revised manuscript was submitted one week after the comments were received.

## References


[1] G.S. Morrison, Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison, Sci. Justice 54 (2014) 245–256, https://doi.org/10.1016/j.scijus.2013.07.004.

[2] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique, Doctoral dissertation, University of Lausanne, 2001.

[3] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang. 20 (2006) 230–275, https://doi.org/10.1016/j.csl.2005.08.001.

[4] Y.A. Solewicz, T. Becker, G. Jardine, S. Gfroerer, Comparison of speaker recognition systems on a real forensic benchmark, in: Proceedings of Odyssey, The Speaker and Language Recognition Workshop, International Speech Communication Association, 2012, pp. 85–91. http://isca-speech.org/archive/odyssey_2012/od12_086.html.

[5] Y.A. Solewicz, G. Jardine, T. Becker, S. Gfroerer, Estimated intra-speaker variability boundaries in forensic speaker recognition casework, in, in: Proceedings of Biometric Technologies in Forensic Science (BTFS), 2013, pp. 30–33.

[6] C. Zhang, G.S. Morrison, E. Enzinger, F. Ochoa, Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices, Speech Commun. 55 (2013) 796–813, https://doi.org/10.1016/j.specom.2013.01.011.

[7] D. van der Vloed, J. Bouten, D. van Leeuwen, NFI-FRITS: A forensic speaker recognition database and some first experiments, in: Proceedings of Odyssey, The Speaker and Language Recognition Workshop, International Speech Communication Association, 2014, pp. 6–13. http://cs.uef.fi/odyssey2014/program/pdfs/21.pdf.

[8] E. Enzinger, Comparison of GMM-UBM and i-vector models under casework conditions: Case 1 revisited, in: Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence, Doctoral dissertation, University of New South Wales, 2016, ch. 4. http://handle.unsw.edu.au/1959.4/55772.

[9] E. Enzinger, G.S. Morrison, F. Ochoa, A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case, Sci. Justice 56 (2016) 42–57, https://doi.org/10.1016/j.scijus.2015.06.005.

[10] D. van der Vloed, Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Communication 85 (2016) 127–130. http://dx.doi.org/10.1016/j.specom.2016.10.001. [errata in: 92 (2017) 23. http://dx.doi.org/10.1016/j.specom.2017.04.005].

[11] E. Enzinger, G.S. Morrison, Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case, Forensic Sci. Int. 277 (2017) 30–40, https://doi.org/10.1016/j.forsciint.2017.05.007.

[12] G.D. da Silva, C.A. Medina, Evaluation of MSR Identity Toolbox under conditions reflecting those of a real forensic case (forensic_eval_01), Speech Commun. 94 (2017) 42–49, https://doi.org/10.1016/j.specom.2017.09.001.

[13] D. van der Vloed M. Jessen S. Gfroerer Experiments with two forensic automatic speaker comparison systems using reference populations that (mis)match the test language, in Proceedings of the Audio Engineering Society Conference on Forensic Audio 2017 paper 2–1.

[14] G.S. Morrison, The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings, Forensic Sci. Int. 283 (2018) e1–e7, https://doi.org/10.1016/j.forsciint.2017.12.024.

[15] C. Zhang, C. Tang, Evaluation of Batvox 3.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Communication 100 (2018) 13–17, https://doi.org/10.1016/j.specom.2018.04.008.

[16] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, A. Alexander, Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 112 (2019) 30–36, https://doi.org/10.1016/j.specom.2019.06.005.


[19] One invitee, who was unable to participate in the original meeting for health reasons, was able to contribute to Stage 2, and is included in the list of authors. One of the supporters was not an original invitee but expressed an interest early in the process.

[17] M. Jessen, J. Bortlík, P. Schwarz, Y.A. Solewicz, Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 111 (2019) 22–28, https://doi.org/10.1016/j.specom.2019.05.002.
[18] M. Jessen, G. Meir, Y.A. Solewicz, Evaluation of Nuance Forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Communication 110 (2019) 101–107, https://doi.org/10.1016/j.specom.2019.04.006.
[19] D. van der Vloed, F. Kelly, A. Alexander, Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA, a forensically realistic database, in: Proceedings of Odyssey, The Speaker and Language Recognition Workshop, 2020, pp. 402–407. http://dx.doi.org/10.21437/Odyssey.2020-57.
[20] G.S. Morrison, Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio, Aust. J. Forensic Sci. 45 (2013) 173–197, https://doi.org/10.1080/00450618.2012.733025.
[21] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality, Sci. Justice 58 (2018) 47–58, https://doi.org/10.1016/j.scijus.2017.06.005.
[22] C. Neumann, M. Ausdemore, Defence against the modern arts: the curse of statistics –Part II: 'Score-based likelihood ratios', Law, Probability and Risk 19 (2020) 21–42, https://doi.org/10.1093/lpr/mgaa006.
[23] C. Neumann, J. Hendricks, M. Ausdemore, Statistical support for conclusions in fingerprint examinations, in: D.L. Banks, K. Kafadar, D.H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC, Boca Raton, FL, 2020, pp. 277–324, https://doi.org/10.1201/9780367527709.
[24] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT, Cambridge, MA, 2012.
[25] B. Found, Deciphering the human condition: The rise of cognitive forensics, Aust. J. Forensic Sci. 47 (2015) 386–401, https://doi.org/10.1080/00450618.2014.965204.
[26] R.D. Stoel, C.E.H. Berger, W. Kerkhoff, E.J.A.T. Mattijssen, E.I. Dror, Minimizing contextual bias in forensic casework, in: K.J. Strom, M.J. Hickman (Eds.), Forensic Science and the Administration of Justice: Critical Issues and Directions, Sage, Thousand Oaks, CA, 2015, pp. 67–86, https://doi.org/10.4135/9781483368740.n5.
[27] National Commission on Forensic Science, Ensuring that forensic analysis is based upon task-relevant information, 2015. https://www.justice.gov/ncfs/file/818196/download.
[28] G. Edmond, A. Towler, B. Growns, G. Ribeiro, B. Found, D. White, K. Ballantyne, R. A. Searston, M.B. Thompson, J.M. Tangen, R.I. Kemp, K. Martire, Thinking forensics: Cognitive science for forensic practitioners, Sci. Justice 57 (2017) 144–154, https://doi.org/10.1016/j.scijus.2016.11.005.
[29] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy, Law Hum Behav. 11 (1987) 167–187, https://doi.org/10.1007/BF01044641.
[30] D.M. Ommen, C.P. Saunders, Building a unified statistical framework for the forensic identification of source problems, Law, Probability and Risk 17 (2018) 179–197, https://doi.org/10.1093/lpr/mgy008.
[31] C.G.G. Aitken, P. Roberts, G. Jackson, Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society, London, UK, 2010.
[32] B. Robertson, G.A. Vignaux, C.E.H. Berger, Interpreting Evidence: Evaluating Forensic Science in the Courtroom, Second ed., Wiley, Chichester, UK, 2016 https://doi.org/10.1002/9781118492475.
[33] G.S. Morrison, E. Enzinger, C. Zhang, Forensic speech science, in: I. Freckelton, H. Selby (Eds.), Expert Evidence, Thomson Reuters, Sydney, Australia, 2018, ch. 99.
[34] G.S. Morrison, E. Enzinger, D. Ramos, J. González-Rodríguez, A. Lozano-Díez, Statistical models in forensic voice comparison, in: D.L. Banks, K. Kafadar, D. H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC, Boca Raton, FL, 2020, pp. 451–497, https://doi.org/10.1201/9780367527709.
[35] J.H.L. Hansen, T. Hasan, Speaker recognition by machines and humans: A tutorial review, IEEE Signal Processing Mag. (2015) 74–99, https://doi.org/10.1109/MSP.2015.2462851.
[36] J.H.L. Hansen, H. Bořil, On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks, Speech Commun. 101 (2018) 94–108, https://doi.org/10.1016/j.specom.2018.05.004.
[37] C. Zhang, G.S. Morrison, E. Enzinger, F. Ochoa, Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices, Speech Commun. 55 (2013) 796–813, https://doi.org/10.1016/j.specom.2013.01.011.
[38] E. Enzinger, G.S. Morrison, The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems, in, in: Proceedings of the Australasian International Conference on Speech Science and Technology, 2012, pp. 137–140.
[39] G.S. Morrison, F. Kelly, A statistical procedure to adjust for time-interval mismatch in forensic voice comparison, Speech Commun. 112 (2019) 15–21, https://doi.org/10.1016/j.specom.2019.07.001.
[40] I.W. Evett, J.S. Buckleton, Statistical analysis of STR data, in: A. Carraredo, B. Brinkmann, W. Bär (Eds.), Advances in Forensic Haemogenetics, Springer-Verlag, Heidelberg, 1996, vol. 6, pp. 79–86.
[41] J. González-Rodríguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-García, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 2104–2115, https://doi.org/10.1109/TASL.2007.902747.
[42] D. Ramos, J. González-Rodríguez, G. Zadora, C.G.G. Aitkin, Information-theoretical assessment of the performance of likelihood ratio computation methods, J. Forensic Sci. 58 (2013) 1503–1518, https://doi.org/10.1111/1556-4029.12233.
[43] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, T. Niemi, Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition, including guidance on the conduct of proficiency testing and collaborative exercises, European Network of Forensic Science Institutes (2015).
[44] G.S. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction, Speech Commun. 85 (2016) 119–126, https://doi.org/10.1016/j.specom.2016.07.006.
[45] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. 276 (2017) 142–153, https://doi.org/10.1016/j.forsciint.2016.03.048.
[46] D.A. van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in: C. Müller (Ed.), Speaker Classification, Springer, Berlin, 2007, vol. 1, pp. 330–335. http://dx.doi.org/10.1007/978-3-540-74200-5_19.
[47] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (2011) 91–98, https://doi.org/10.1016/j.scijus.2011.03.002.